Deep multi-modal data analysis and fusion for robust scene understanding in CAVs

1st Andreas Papandreou*
Department of Electrical and Computer Engineering
University of Patras
University Campus, 26504 Rio, Greece
apapandreou@ece.upatras.gr

2nd Andreas Kloukiniotis*
Department of Electrical and Computer Engineering
University of Patras
University Campus, 26504 Rio, Greece
kloukiniotisandreas@ece.upatras.gr

3rd Aris Lalos
I.S.I. - Industrial Systems Institute
Patras Science Park building
Platani, PATRAS, Greece
lalos@isi.gr

4th Konstantinos Moustakas
Department of Electrical and Computer Engineering
University of Patras
University Campus, 26504 Rio, Greece
moustakas@ece.upatras.gr

2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP) | 978-1-6654-3288-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/MMSP53017.2021.9733604

***Abstract*—Deep learning (DL) tends to be the integral part of Autonomous Vehicles (AVs). Therefore the development of scene analysis modules that are robust to various vulnerabilities such as adversarial inputs or cyber-attacks is becoming an imperative need for the future AV perception systems. In this paper, we deal with this issue by exploring the recent progress in Artificial Intelligence (AI) and Machine Learning (ML) to provide holistic situational awareness and eliminate the effect of the previous attacks on the scene analysis modules. We propose novel multi-modal approaches against which achieve robustness to adversarial attacks, by appropriately modifying the analysis Neural networks and by utilizing late fusion methods. More specifically, we propose a holistic approach by adding new layers to a 2D segmentation DL model enhancing its robustness to adversarial noise. Then, a novel late fusion technique has been applied, by extracting direct features from the 3D space and project them into the 2D segmented space for identifying inconsistencies. Extensive evaluation studies using the KITTI odometry dataset provide promising performance results under various types of noise.**

***Index Terms*—autonomous vehicles, multi-modal scene analysis, adversarial attacks**

I. Introduction

AVs are considered as an important component of connected intelligent transportation systems, enhancing travel security, fuel economy and the travel experience of road users. One of the most essential operations executed at the AVs, to enable the aforementioned benefits, is the perception and understanding of dynamic and complex environments from multi-modal sensor data. There are roughly three approaches [1], utilized in deep multi-modal object detection, called early, middle and late fusion while late fusion is widely adopted due to the modularity benefits that it offers.

*The first two authors had equal contribution.

Despite their great success, DL techniques introduce formidable challenges in dealing with carefully crafted adversarial perturbations [2]. Cyber-attacks have damaging effects on an industry like the Cooperative Connected and Automated Mobility (CCAM). From the least important to the worst ones, one can mention for example the damage in the reputation of vehicle manufacturers, the increased denial of customers to adopt CCAM, the loss of working hours (having a direct impact on the European GDP), material damages, increased environmental pollution and ultimately the great danger for human lives, either they are drivers, passengers or pedestrians. Thus there is an increasing interest in both academia and industry to proactively address modern vehicle cybersecurity challenges applying advanced AI and ML techniques, and seeking methods to mitigate associated safety risks.

Within this work we will robustify the performance of multi-modal approaches by utilizing early pre-processing and late fusion methods. More specifically, we design and implement new layers that are added in a unified manner to the 2D/3D analysis networks increasing their robustness in adversarial noise. Then we propose a novel late fusion method that initially extracts useful features from the 3D point clouds and project them into the 2D segmented image for identifying inconsistencies. Extensive evaluations utilizing the KITTI odometry dataset highlight the benefits of the proposed methods under various scenarios with different types of noise.

II. Countering adversarial attacks

The direction of this paper focus mainly on providing a robust fusion scheme against adversarial attacks to the camera sensor. The overall architecture could be divided into two different parts. The former is dedicated to the pre-processing and the analysis of the captures 2D and 3D scenes, while the

latter is based on the co-registration and processing of data from multiple sources located at different strategic points on the vehicle. In the next sections, after a short introduction to adversarial attacks, we describe the proposed DL solution. The scene analysis is performed using a robust 2D segmentation model aiming to alleviate the adversarial noise and recognize the environment. A 3D object detection model is also utilized, that process raw point clouds coming from the lidar. Finally, a robust decision is taken by fusing the outputs of the two analysis modules. The overall pipeline is presented in Fig. 1.

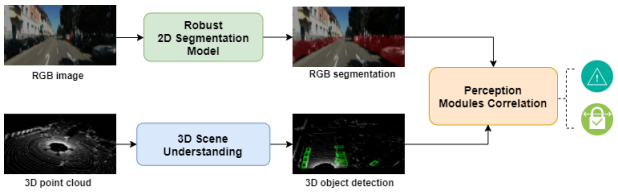

Fig. 1. Proposed architecture: Image and point cloud inputs are being processed by the 2D segmentation and 3D object detection modules accordingly. By correlating the two output at the decision stage, an indication is shown to the user for a dangerous or safe situation.

A. Adversarial Attacks on Autonomous Vehicles:

Deep neural networks have being widely utilized in various autonomous driving scenarios, offering various benefits but posing at the same time and great concerns for the security and integrity of the applications. Adversaries can alter original inputs with perturbations, which may be imperceptible to the human eye, but can force a trained model to produce incorrect outputs. Szegedy et al. [2] first discovered that state-of-art deep neural networks are susceptible to adversarial attacks. Adversarial examples seem to occur from the extreme non linearity of deep neural networks. Studies on adversarial attacks have developed attacks for image classification models [3], [4], used for multiple vision tasks such as object detection [5], [6], object tracking [7], and semantic segmentation [6]. A comprehensive study toward adversarial robustness was presented by Arnab et al [8]. They evaluated the robustness of popular DCNN models used for segmentation tasks against adversarial attacks and concluded that the accuracy of the models seriously decreased after the original image has been perturbed. Adversaries have also being discovered to be mistaken by DCNNs for traffic legitimate traffic signs [9]. The aforementioned methods were generating attacks aiming to fool the perception systems based on the camera sensor. However, Xiang et al [10] proposed attacking methods to generate adversarial point clouds. The latter managed to fool a widely used neural network for point cloud processing, achieving a high error rate.

B. Adversarial Defense using Denoising Methods:

Image restoration techniques could be considered as a denoising block that will precede the execution of the DL model, to eliminate the impact of the adversaries to the perception system output. We evaluated the ability of various state of the art methods that have been either proposed for image restoration or for adversarial noise removal in mitigating adversarial attacks in scene analysis operations. Both of them manage to restore the attacked image. However, there is a limitation on methods trained on adversarial noise only, aiming to remove the perturbation of the adversaries. The majority of them can only work on a limited range of resolutions, making them impractical to applications related to Connected AVs (CAVs). Hence, we concentrate only on methods that address the topic of the image restoration problem. The predominant approaches, from the previous category have been working mostly with Gaussian noise and similar random noise models to corrupt images.

Starting with Zhang et al [11], he pointed out that residual learning and batch normalization can benefit each other. Their integration was effective in speeding up the training and boosting denoising performance. Although a trained feed-forward denoising convolutional neural network is able to handle compression and interpolation errors, the trained model under a given noise variance (e.g., σ) is not suitable for other noise variances. In a noise agnostic case where the noise level σ is unknown, the denoising method should let the user define a trade-off between noise suppression and texture protection. The user-directed approach FFDNet was introduced by [12]. The proposed method takes as input the noise level which makes it flexible to different noises. In 2019, it was introduced Adaptive Feature Modification Layers(AdaFm) [13] in a step toward handling continual modulation of restoration levels. AdaFm enables consecutive modulation of the restoration strength at a considerable low computation cost. At first a standard restoration CNN is trained for the start level, and then AdaFM layers are inserted to optimize it to the end level. After the training stage, CNN parameters are being fixed. The filters of AdaFM layers are interpolated according to the testing restoration level. By using a controlling coefficient, the CNN is able to manipulate the restoration effects. Finally, AdaFM has been integrated into our pipeline, in order a robust 2D segmentation model. More details about the integration are explained in a further section.

III. Proposed Architecture for Robust CAVs

The proposed architecture for robust CAVs is presented here and each module is described separately.

A. 2D Semantic Segmentation of the Scene:

Semantic segmentation is a fundamental problem in computer vision and is necessary for higher-level tasks such as scene understanding or object detection. Although, it is rather a complex problem due to the complicated object boundaries and the large number of classes that the model needs to distinguish. The success of deep convolutional neural networks could not leave unaffected the segmentation approaches. Depth of the CNN as shown by [15], [16] is crucial for providing rich features but the accuracy decreases due to the higher complexity of the model. ResNet [17] is a state of the art approach that addresses this problem. Several ResNet-like methods have

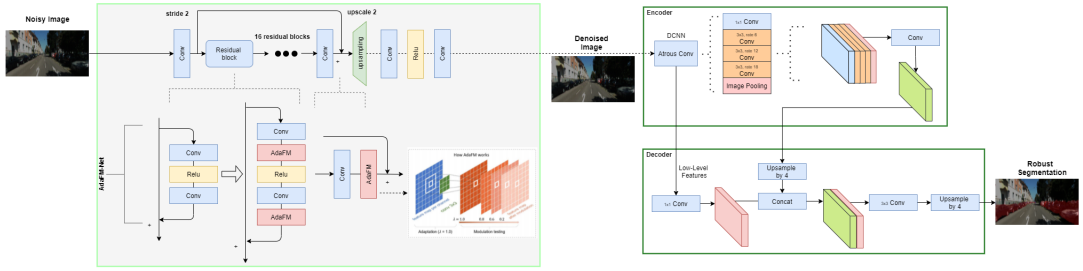

Fig. 2. Fusion model for generating a robust 2D segmentation model, from left to right: the AdaFM denoising model [13] and the Deeplabv3 segmentation model [14]

been published [18] improving the accuracy. To decode more efficient the global context information and perform pixel-level prediction, Zhao et al. [19] introduced a pyramid scene parsing network (PSPNet). Features of several pyramid scales, that combine both local and global context information, are fused by applying a pyramid parsing module in the last convolutional layer of a CNN. The final per-pixel prediction is obtained by a convolution layer. Inspired by PSPNet [19], Deeplab [14] revisited the Atrous Spatial Pyramid Pooling (ASPP) [20] by experimenting with cascading and parallel application of dilated convolutions. This allows them to improve upon their previous work [14] while achieving comparable results to PSPNet [19]. Deeplab v3 is the model that we also use in the proposed architecture due to its efficiency in comparison to other relevant state of the art approaches. As denoted by the [21] it performs multi-scale processing and should be preferred in safety-critical applications due to its inherent robustness against adversarial attacks.

B. Denoising Model Integration:

Our goal is to investigate how image denoising can enhance high-level vision applications, including semantic segmentation tasks, especially in CAVs. In the proposed end-to-end architecture that has been implemented, see Fig. 2, the denoising block is added prior to the segmentation model. We focus on ensemble learning so as to improve the performance of scene analysis operations in CAVs. Both models were trained separately using data from the KITTI dataset [22]. The first part of the proposed integrated model is following the structure of the AdaFM [13] technique. The latter forwards the denoised image to the segmentation model and it follows the structure of Deeplab [14]. As a result, the proposed model significantly increases the robustness of the 2D segmentation model, alleviating the effect of various adversarial perturbations.

C. 3D Object Detection of the Scene:

Despite achieving state-of-art results, camera-based approaches are heavily influenced by their physical limitations. In contrast lidars are not susceptible to environmental factors. In our case, by assuming that only the camera modality can be attacked, mapping features that are extracted from the captured point clouds to the image space is proposed as a solution to the attack. In this section, a short review of the DL frameworks that have been applied to 3D data is all presented for the sake of self-completeness. Many researchers have developed efficient representations to detect and localize objects in point clouds. Point clouds lack a specific structure, and researchers trying to exploit them by using approaches that are analysing the 3D space in the form of voxel grids [23]–[25], raw point clouds [26], [27] or by processing them in 2D feature maps acquired by projection.

In Voxel-Based models, data are separated into uniform grids with fixed dimensions to represent the distribution of the data in 3D space. Typically, the size of the grid is established according to the resolution of the data. The main advantage of the representation based on voxels is that it can encode the 3D shape and the viewpoint information by classifying the occupied voxels into several types, such as visible, occluded, or self-occluded. Besides, 3D convolution and pooling operations can be directly applied in voxel grids. 3D ShapeNet proposed by Wu et al. [23], is the pioneer in exploiting 3D volumetric data using a convolutional deep belief network. VoxNet is proposed in [24] and conducts 3D object recognition employing 3D convolution filters based on volumetric data design.

In order for the data to be processed with classic 2D convolutional layers, several methods project 3D point clouds in 2D grid-based features maps. Spherical space [28], camera-plane map (CPM) and bird's eye view (BEV) [29] space are the most dominant approaches. The spherical map is obtained by projecting the point cloud onto a sphere. This is a dense and compact way of representing the point cloud, but it stills differs from images and the fusion is not straightforward. CPM can be directly fused with camera images at any stages of the CNN, but lidar resolution is not as dense as images, which makes an upsampling [30] necessary. BEV is suitable for object localization because it directly provides the positions of the objects on the ground, maintaining objects length and width.

Regarding point cloud-based models, significant impact has the deployment of PointNet [26] and later the PointNet++ [27]. PointNet [26], as a pioneer in consuming 3D point clouds directly for deep models, learns the spatial feature of each point independently via MLP layers and then accumulates their features by max pooling. The point clouds are given as input

directly to the PointNet, which predicts the per-point label or per-object label. In PointNet, a spatial transform network and a symmetric function are designed to improve the data invariance to permutation. PointRCNN [31] which will be integrated into our pipeline, achieves state-of-art results in a two-stage 3D object detection framework. The first stage segments foreground points and generates a small number of bounding box proposals from the segmented points simultaneously, while the second stage conducts canonical 3D box refinement. An extension of [31] from the same authors is Part-A^2 [32] which is a part-aware and aggregation neural network. Finally, an interesting approach is PV-RCNN [33], aiming to take advantage of efficient learning and high-quality proposals of the 3D voxel CNN and the flexible receptive fields of the PointNet-based networks.

IV. Fusion Strategy

There is a variety of strategies aiming to increase the overall performance, by fusing multiple modalities. We can divide the previous types into three categories: Early fusion, Late fusion and Deep fusion. Taking into consideration the first category, modalities are combined at the beginning of the process and extract the shared information on data, by jointly processing the raw data measurements acquired by different sensors. Late fusion performs the synthesis of valuable information at the final stage of feature extraction, where fusion occurs. Finally, a more general fusion scheme refers to Deep fusion. By exploiting the capability of DL to discover high-level data representation, the Deep fusion can effectively find the joint data representation by combining the features extracted at the intermediate layers of deep neural networks. In our case, late fusion has been applied.

A. Point Projection

For projecting lidar points to the image plane, we are using the calibration data provided by the KITTI benchmark [22]. lidar data have been captured by a Velodyne HDL-64E S2 sensor and each point is stored with its (x, y, z) coordinate and an additional reflectance value (r). The number of points per scan on average for each file/frame is $\sim 120,000$ 3D points. The rigid body transformation from Velodyne coordinates to camera coordinates are given in detail at [22] and are expressed by:

$$T_{velo}^{cam} = \begin{pmatrix} R_{velo}^{cam} & t_{velo}^{cam} \\ 0 & 1 \end{pmatrix}$$

where $R_{velo}^{cam} \in R^{3\times3}$ is the rotation matrix and $t_{velo}^{cam} \in R^{1\times3}$ the translation vector from Velodyne to the camera coordinate system. The projection of a 3D point $\mathbf{x} = (x, y, z, 1)^T$ in the lidar coordinates system gets projected to a point in the camera plane $\mathbf{y} = (u, v, 1)^T$ according to:

$$\mathbf{y} = T_{velo}^{cam}\mathbf{x}$$

B. Perception Modules Correlation

Our purpose is the situational awareness improvement and the mitigation of cyber-attacks against computer vision systems, especially on the camera sensor. Firstly, we apply 2D semantic segmentation to the input image to recognize the objects in the scene. In a safe situation, the majority of the objects will be detected, whereas in an attacked case, some of them will be disappeared from the perception engine. In a parallel module, the 3D object detection model will be implemented, which has been trained to identify only the moving objects, which are mainly vehicles, pedestrians and cyclists. As soon as the coordinates of the 3D bounding boxes have been obtained, they are being projected to the image plane. The correlation is applied in the next step in order to fuse the 2D segmentation and 3D object detection outputs, as Fig. 3 illustrates. In the latter, with a red mask and green bounding boxes, the 2D segmentation and 3D detection results are shown accordingly.

Fig. 3. From top to bottom: 2D segmentation output and 3D detected objects projected to the image plane

To correlate the two outputs, we isolate the region of the projected 3D bounding box to the image. We consider that the isolated region belongs to a specific class (vehicle, pedestrian, cyclist). We isolate respectively the same region from the segmentation mask. Finally, we compare the two outputs in order to estimate the overlap between the two detected regions. If an object has been detected by both of the modalities, the overlap should be high enough. Structural similarity index measure (SSIM) was used for the comparison and the threshold for a safe situation should be above 0.6 for the majority of the detected objects.

V. EXPERIMENTS & RESULTS

The proposed algorithms were evaluated using the KITTI dataset [22]. The results between the robust and the original 2D segmentation model are being presented in Table I. There is a variety of approaches [34] aiming to generate strong adversarial attacks with a high error rate in order to manipulate the vehicle's behaviour. We tested the models with three generic adversarial attacks, BIM [35], FGSM [3], and PGD [36] attack. As we can observe from Table I, the robust model achieves great IoU scores even when the magnitude of perturbation is large. The previous perturbation refers to a hyper-parameter governing the distance between adversarial and original image. The smaller the magnitude of the perturbation, the less imperceptible will be the attack to the human eye.

We evaluated the previous attacks setting the perturbation to each value from {2,4,8} on the scale of [0-255]. The iterations for PGD iterative attack is set to 10 and the step-size is equal to 1, so each pixel could be changed by one each iteration. For instance, observing Table I, the original model fails to operate after the PGD attack, whereas the robust model achieves 72% IoU score. An example is illustrated also in Fig. 4 in which the robust segmentation model manages to restore the attacked image.

TABLE I
IOU% RESULTS OF SEGMENTATION MODELS FOR ADVERSARIAL ATTACKS WITH LEVELS OF PERTURBATIONS 2/4/8 FOR A GIVEN IMAGE

Model	BIM	FGSM	PGD
Robust	0.42 / 0.1 / 0.03	0.75 / 0.71 / 0.66	0.74 / 0.72 / 0.72
Original	0.13 / 0.02 / 0.01	0.67 / 0.51 / 0.21	0.69 / 0.53 / 0.17

Fig. 4. From top to bottom: original segmented image, attacked segmented image, robust segmented image (with the denoiser)

The results of PointRCNN [31] which was used as the object detection module was validated on KITTI dataset [22]. Table II presents the 3D detection performance of moderate difficulty on the validation set of KITTI dataset. PointRCNN obtains 78.70% recall, given an IoU threshold at 0.5 on the moderate difficulty for the car class, 54.41% for pedestrian and 72.11% for cyclist.

TABLE II
KITTI [22] RESULTS FROM OPENPCDET [37] FRAMEWORK

Model	Car	Pedestrian	Cyclist
PointRCNN	78.70	54.41	72.11
$Part - A^2$-Free	78.72	65.99	74.29
$Part - A^2$-Anchor	79.40	60.05	69.90
PV-RCNN	83.61	57.90	70.47

Overall, by fusing the outputs of multiple perception modules, it is possible to provide improved situational awareness. Understanding the autonomous vehicle's state at any time is critical to identifying potential threats and generate secure transportation systems. In Fig. 5, a scenario without adversarial attack on the camera sensor is illustrated. On the left, the 3D object detection output from the lidar data is shown. With green and blue bounding boxes, the vehicles and pedestrians are shown accordingly. On the bottom image, the 2D robust segmentation result is presented, with coloured masks. Therefore, by correlating the previous outputs, we raise a safe situation, considering that the majority of the existed objects in the scene has been detected in both sensors. Thus, the final decision of the perception engine is coming only from the camera sensor, as we can observe from the top side of Fig. 5. On the other hand, Fig. 6 indicates an attacked scenario, in which an external attacker has added adversarial noise to the camera image. As a result, the 2D segmentation module fails to recognize the scene and raises an alert to the user. With orange bounding boxes in the bottom image, the hidden objects from the camera sensor are presented. Hence, the final decision of the perception engine is coming only from lidar, by projecting the 3D outputs to the image plane, as we can observe from the top side of Fig. 6.

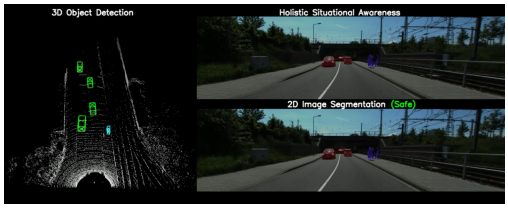

Fig. 5. From left to right and bottom to top: 3D object detection output, 2D robust segmentation output and the fused result, indicating a safe situation

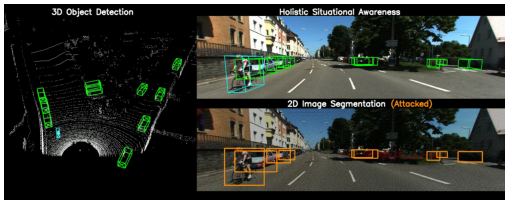

Fig. 6. From left to right and bottom to top: 3D object detection output, 2D robust segmentation output and the fused result, indicating an attacked situation

VI. CONCLUSIONS

Our goal is to achieve contextual and situational awareness, by fusing different data sources of information to facilitate the decision-making process. Overall, we have presented a robust fusion scheme for providing situational awareness to the driver. We have integrated a denoising module atop an image segmentation CNN to improve its robustness to adversarial inputs. As a second layer of defense, the result of an object detection module on lidar data is correlated with the image segmentation output. Thus, anomaly detection on the camera sensor could be detected, leading to more secure perception systems for AVs.

In future work, we will investigate scenarios where the lidar is attacked. One mitigation strategy for such scenarios could be to decide whether the lidar is attacked or not by checking the consistency with the output of the image segmentation before and after the denoising. In particular, if the camera has not been attacked, meaning that scene segmentation will not

differ before and after the denoising, and the scene analysis of image and lidar data disagrees then lidar data probably have been modified. Different levels of trust should be also defined so as to put different weights on each separate sensor data. As such, by correlating the outputs of different perception modules with additional sensor readings, it is possible to provide improved situational awareness. Another action point could be the definition of a more complex strategy for fusing the multiple modalities, aiming to improve the whole system performance.

ACKNOWLEDGMENT

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No.833611 (CARAMEL).

REFERENCES

[1] Eduardo Arnold, Omar Y. Al-Jarrah, Mehrdad Dianati, Saber Fallah, David Oxtoby, and Alex Mouzakitis. A survey on 3d object detection methods for autonomous driving applications. *IEEE Transactions on Intelligent Transportation Systems*, 20(10):3782–3795, 2019.

[2] Christian Szegedy, W. Zaremba, Ilya Sutskever, Joan Bruna, D. Erhan, Ian J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2014.

[3] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2015.

[4] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Discovering adversarial examples with momentum. *CoRR*, abs/1710.06081, 2017.

[5] Xingxing Wei, Siyuan Liang, Xiaochun Cao, and Jun Zhu. Transferable adversarial attacks for image and video object detection. *CoRR*, abs/1811.12641, 2018.

[6] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan L. Yuille. Adversarial examples for semantic segmentation and object detection. *CoRR*, abs/1703.08603, 2017.

[7] Bin Yan, Dong Wang, Huchuan Lu, and Xiaoyun Yang. Cooling-shrinking attack: Blinding the tracker with imperceptible noises, 2020.

[8] A. Arnab, O. Miksik, and P. H. S. Torr. On the robustness of semantic segmentation models to adversarial attacks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 888–897, 2018.

[9] Y. Li, X. Xu, J. Xiao, S. Li, and H. T. Shen. Adaptive square attack: Fooling autonomous cars with adversarial traffic signs. *IEEE Internet of Things Journal*, pages 1–1, 2020.

[10] Chong Xiang, Charles R. Qi, and Bo Li. Generating 3d adversarial point clouds. *CoRR*, abs/1809.07016, 2018.

[11] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.

[12] K. Zhang, W. Zuo, and L. Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018.

[13] Jingwen He, Chao Dong, and Yu Qiao. Modulating image restoration with continual levels via adaptive feature modification layers. *CoRR*, abs/1904.08118, 2019.

[14] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017.

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[16] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[18] Zifeng Wu, Chunhua Shen, and Anton van den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition, 2016.

[19] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network, 2017.

[20] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, 2017.

[21] Anurag Arnab, Ondrej Miksik, and Philip H. S. Torr. On the robustness of semantic segmentation models to adversarial attacks, 2018.

[22] Jun Xie, Martin Kiefel, Ming-Ting Sun, and Andreas Geiger. Semantic instance annotation of street scenes by 3d to 2d label transfer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[23] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes, 2015.

[24] D. Maturana and S. Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 922–928, 2015.

[25] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions, 2017.

[26] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation, 2017.

[27] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space, 2017.

[28] Bichen Wu, Alvin Wan, Xiangyu Yue, and Kurt Keutzer. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud, 2017.

[29] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving, 2017.

[30] Andreas Pfeuffer and Klaus Dietmayer. Optimal sensor data fusion architecture for object detection in adverse weather conditions, 2018.

[31] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud, 2019.

[32] Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network, 2020.

[33] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection, 2019.

[34] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay. Adversarial attacks and defences: A survey. *ArXiv*, abs/1810.00069, 2018.

[35] A. Kurakin, Ian J. Goodfellow, and S. Bengio. Adversarial examples in the physical world. *ArXiv*, abs/1607.02533, 2017.

[36] R. S. Zimmermann. Comment on "adv-bnn: Improved adversarial defense through robust bayesian neural network". *ArXiv*, abs/1907.00895, 2019.

[37] OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds. urlhttps://github.com/open-mmlab/OpenPCDet, 2020.